NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

A METHOD FOR CLASSIFICATION OF INCOMPLETE NETWORKS: TRAINING THE MODEL WITH COMPLETE AND INCOMPLETE INFORMATION

by

Carolyne Vu

March 2019

Thesis Advisor:	Ruriko Yoshida
Second Reader:	David L. Alderson Jr.

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE	*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** March 2019	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE A METHOD FOR CLASSIFICATION OF INCOMPLETE NETWORKS: TRAINING THE MODEL WITH COMPLETE AND INCOMPLETE INFORMATION	**5. FUNDING NUMBERS**
6. AUTHOR(S) Carolyne Vu	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The rise of accessible real-world data creates a growing interest in effective methods for accurate classification, especially for networks with incomplete information. The intelligence community requires an understanding of a network before the team can develop a strategy to combat the adversary. These problems are typically time-sensitive; however, gathering complete and actionable intelligence is a challenging mission. An adversary's actions are secretive in nature. Crucial information is deliberately concealed. Intentionally dubious information creates problematic noise. Therefore, if an observed incomplete network can be classified as-is without delay, the network can be properly analyzed for a strategy to be devised and acted upon earlier. This thesis considers a machine learning technique for classification of incomplete networks. We examine the effects of training the model with complete and incomplete information. Observed network data and their structural features are classified into technological, social, information, and biological categories using supervised learning methods.

14. SUBJECT TERMS network analysis, incomplete information, network classification, graph theory	**15. NUMBER OF PAGES** 61
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

THIS PAGE INTENTIONALLY LEFT BLANK

A METHOD FOR CLASSIFICATION OF INCOMPLETE NETWORKS: TRAINING THE MODEL WITH COMPLETE AND INCOMPLETE INFORMATION

Carolyne Vu
Lieutenant, United States Navy
BS, U.S. Naval Academy, 2012

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

**NAVAL POSTGRADUATE SCHOOL
March 2019**

Approved by: Ruriko Yoshida
Advisor

David L. Alderson Jr.
Second Reader

W. Matthew Carlyle
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The rise of accessible real-world data creates a growing interest in effective methods for accurate classification, especially for networks with incomplete information. The intelligence community requires an understanding of a network before the team can develop a strategy to combat the adversary. These problems are typically time-sensitive; however, gathering complete and actionable intelligence is a challenging mission. An adversary's actions are secretive in nature. Crucial information is deliberately concealed. Intentionally dubious information creates problematic noise. Therefore, if an observed incomplete network can be classified as-is without delay, the network can be properly analyzed for a strategy to be devised and acted upon earlier. This thesis considers a machine learning technique for classification of incomplete networks. We examine the effects of training the model with complete and incomplete information. Observed network data and their structural features are classified into technological, social, information, and biological categories using supervised learning methods.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

CS	Computer Science
DoD	Department of Defense
MAE	Mechanical and Aerospace Engineering
ML	Machine Learning
NN	Neural Network
NPS	Naval Postgraduate School
OA	Operations Analysis
RF	Random Forest
SHAP	Shapley Additive Explanations
SVM	Support Vector Machine
UAV	Unmanned Autonomous Vehicle

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Demand for effective methods of analyzing networks has emerged with the growth of accessible data, particularly for incomplete networks. Even as means for data collection advance, incomplete information remains a reality for numerous reasons. Data can be obscured by excessive noise. Surveys for information typically contain some non-respondents. In other cases, simple inaccessibility restricts observation. Also, for illicit groups, we are confronted with attempts to conceal important elements or their propagation of false information. In the real-world, it is difficult to determine when the observed network is both accurate and complete.

In this research, we consider a method for classification of incomplete networks. We classify real-world networks into technological, social, information, and biological categories by their structural features using supervised learning techniques. In contrast to the current method of training models with only complete information, we examine the effects of training our classification model with both complete and incomplete network information. This technique enables our model to learn how to recognize and classify other incomplete networks.

The representation of incomplete networks at various stages of completeness allows the machine to examine the nuances of incomplete networks. By allowing the machine to study incomplete networks, its ability to recognize and classify other incomplete networks improves drastically. Our method requires minimal computational effort and can accomplish an efficient classification. The results strongly confirm the effectiveness of training a classification model with incomplete network information.

The foundation established in this thesis allows for an enhanced understanding of incomplete networks. Opportunities for follow-on research extend to incorporation of this classification model into practical implementation and exploration of other machine learning techniques.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

Many thanks to my family and friends for all of their support throughout my time at the Naval Postgraduate School (NPS). In particular, I must thank my wife, Elizabeth, for her steady love and confidence in me. I could not have accomplished this without you.

To my peers, it has been a joy learning beside you. To the faculty and staff, thank you for your instruction. It has been my honor learning in your classrooms and to be a member of this esteemed department.

Finally, to my advisors, thank you for all of the time, guidance and enthusiasm you dedicated to me throughout this process.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

1.1 Motivation

Numerous real-world problems and systems can be represented by networks. Varying from social relationships to biochemistry, networks exist in many different forms. Personal or organizational interactions are captured in network representations (Barabasi 2016, sec.1.5). Communication devices, connected through wired or wireless means, can be mapped by technological networks (Barabasi 2016, sec.1.5). Networks can also capture military coordination necessary to an operation to assist in decision-making (Barabasi 2016, sec.1.5). Network analysis provides a discipline shared by many different professional fields. Biologists and intelligence analysts alike often characterize their system, extract information from potentially incomplete data, and develop an understanding of their system through analysis (Barabasi 2016, sec.1.4).

The initial stages of any network analysis includes categorization by class according to shared characteristics Newman (2010, p.13). This is known as classification. From an accurate classification, similar methods of analysis can be applied to networks belonging in the same class.

The rise of accessible real-world data creates a growing interest in effective methods for accurate network classification, especially for networks with incomplete information. The principle challenge of analyzing real-world observed networks is its propensity to contain dubious or incomplete data. For example, criminal networks are "inevitably incomplete" given their elusive and dynamic operational nature (Sparrow 1991). Illicit groups might propagate false information to conceal true intentions. Even naturally occurring networks could have elements that are simply unobserved. Also, for some networks, due to their nature or size, it can be difficult to ascertain when the observed network is considered complete.

Unfortunately, very little research has been completed to study the effects of incomplete data on network structure (Sparrow 1991). Most techniques for handling incomplete networks involve data imputation, the process of estimating unknown data from the observed data, which might incur unknown consequences to a network's true structure and ultimately affect classification (Little and Rubin 2014, p.20).

An intelligence community's assessment of enemy organizations requires accurate classification of the observed network before the intelligence team can develop a strategy for combating the adversary. Problems are typically time-sensitive; however, gathering this complete and actionable intelligence is a challenging mission that could span years. An adversary's actions are secretive in nature, making it extremely difficult to collect a complete observation of the network. Crucial information is deliberately concealed. Intentionally dubious information might create problematic noise or false imputations. Thus, if an observed incomplete network can be classified as-is without delay, the network can be properly analyzed for a strategy to be devised and acted upon earlier.

With a method to accurately classify an incomplete network, techniques of imputation can be reserved for post-classification. This allows for the estimation to be tailored accordingly by network class in an effort to maintain the network's true structure. These techniques could provide the intelligence team with a reasonable evaluation of an enemy's prospective associations or activities.

A method for classifying incomplete networks has a wide range of potential applications, from social network analysis, to epidemiology, and political campaigning. Incomplete network classification without imputation creates the possibility for new approaches to network analysis.

1.2 Objective

In this research, we consider a method for classification of incomplete networks. We examine the effects of training the classification model with complete and incomplete information. Observed network data and their network features are classified into technological, social, information, and biological categories using supervised learning methods. This comparative analysis contributes to a better understanding of network characteristics for classification.

1.3 Structure

This thesis is organized as follows. Chapter 2 reviews relevant literature of incomplete networks. Chapter 3 describes the network data used, the data preparation process, and the model. Chapter 4 presents the results and deductions from our analysis. Chapter 5 summarizes key findings and recommends areas for continuing research.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
Background

This chapter begins by describing frequently studied networks, separating them into four common classes: technological, social, information, and biological networks. Next, Section 2.2 reviews current literature for network classification and incomplete network studies that we build upon in our own study. Section 2.3 of this chapter provides basic definitions of the terminology used in this thesis.

2.1 General Network Classes

Networks can be used to model a variety of systems. Classifying them into distinguished categories allows for treating networks in a category with common methods of analysis. In this study, we follow the categorization of networks by Newman (2010, p.13) into four general classes.

2.1.1 Technological Networks

Technological networks are used to model physical infrastructure systems fundamental to modern society (Newman 2010, p.13). The Internet, as a global network of connections between devices, transportation networks, power grids, telephone and delivery networks, are included in this category, though they are not all are examined in this study. Technological networks, for example, can have nodes representing airports and the edges representing connections between those airports.

2.1.2 Social Networks

Social networks model people or groups in some form of social interaction connecting them. In popular terms, social networks commonly refer to online network systems such as Facebook or LinkedIn (Barabasi 2016, sec.1.3). However, the study of social networks also includes email interaction, professional collaboration, and familial ancestry (Newman 2010, p.30). For social networks, nodes represent individuals, and edges are their connections.

2.1.3 Information Networks

Information networks represent shared data connections. Closely resembling social networks and some technological networks, information networks actually represent the content occurring over those other networks. Examples include information flow over the World Wide Web, citations, recommendations, and information distribution in the form of sharing others' posts (Newman 2010, p.51). Information networks, for example, can have nodes representing movies and edges representing related recommendations.

2.1.4 Biological Networks

Biological networks are interactions between biological elements (Newman 2010, p.64). Common types of biological networks are used to represent biochemical interactions, neurological systems, and relationships in an ecosystem. Biological networks, for example, can have nodes representing genes and edges representing their interactions.

2.2 Literature Review

Our current "era of big data" is emerging from an increased ability to collect and share data. (Murphy 2012, p.1). The sharp growth in data size and accessibility requires appropriate methods for processing and analyzing that big data. A machine learning (ML) approach to analysis is leveraged in our study.

2.2.1 Machine Learning

ML can be described as a method capable of automated pattern detection to make a prediction or decision under conditions of uncertainty (Murphy 2012, p.1). If leveraged properly, the machine will be able to discover and learn from nuanced patterns undetected by human analysis. Two types of ML are generally used, supervised and unsupervised, also referred to as predictive and descriptive (Murphy 2012, p.2).

The unsupervised (or descriptive) approach is given only inputs with an objective to discover patterns in the data (Murphy 2012, p.9). "This is sometimes called knowledge discovery" (Murphy 2012, p.9). Unsupervised approaches have the flexibility to handle less straightforward problems.

Supervised learning has become a central method of ML. In the supervised (or predictive approach), the objective is to learn from analyzing the relationship between inputs and outputs, given a training set of labeled input and output data (Murphy 2012, p.3). After learning from the patterns observed, the machine should be able to receive an input whose output remains unknown and make a prediction of the output. When outputs are categorical, this is known as classification (Murphy 2012, p.2). When class labels are greater than two and mutually exclusive, it is considered a multiclass classification (Murphy 2012, p.3). For this thesis, the term “classification” refers to a multiclass classification with a singular output.

This thesis employs a supervised learning approach to analyze and classify incomplete networks using a classification model. We accomplish this using ensembles of ML trees.

2.2.2 Random Forest

Classification models have accomplished significant improvements with ML methods. One of the most popular models is decision trees. Decision trees achieve classification estimates by sorting instances down a tree process. Each step of the tree "specifies a test of some attribute of the instance, and each branch descending from that [step] corresponds to one of the possible values for this attribute" (Mitchell 1997, p.53).

"An ensemble of decision trees is called a decision forest" (Alpaydin 2014, p.234). By combining predictions from many decision trees, the overall forest’s accuracy significantly improves and reduces variance (Alpaydin 2014, p.235). This is the foundation of the random forest (RF) model. Repeatedly generating multiple trees to random subsets of the input data forms the forest. This method was formally introduced by Breiman (2001).

Breiman (2001) discovered the benefits of using a RF model include:

- **Accuracy** – RF maintains a competitive accuracy rate among decision tree models.
- **Robustness** – RF is resilient to outliers and noise.
- **Reduced Variance** – The randomness helps reduce issues of variance.
- **Efficiency** – RF models can be fast, especially compared to other ensemble tree models.
- **Model Self-Evaluation** – RF models can provide their own "internal estimates of

error, strength, correlation and variable importance" (Breiman 2001).
- **Simplicity** – RF is a simple model to use.

While RF models remain one of the most commonly used ML methods, there is no universal selection for all classification problems (Murphy 2012, p.551). Using a RF classification model may prove to work well under certain conditions and poorly in others. Some potential drawbacks to using RF include possible overfitting for datasets with high levels of noise, results from the black-box style model can be difficult to interpret, and training speeds are slower with large datasets (Louppe 2014). An understanding of the benefits and drawbacks to the RF method is necessary when constructing a RF model.

2.2.3 Study of Network Classification

The initial phase of any network analysis begins with classifying the network. Geng et al. (2012) discusses the following kernel methods popular for network classification.

- **Random Walk** – similarity measured by common random walks
- **Shortest Path** – similarity measured by common shortest paths
- **Cyclic Pattern** – similarity measured by common cycles
- **Subtree** – similarity measured by common subtrees
- **Graphlet and Subgraph** – similarity measured by similar subgraphs or graphlets

Geng et al. (2012) proposed an alternative approach to kernel methods. It is commonly understood that networks of a class will have similar characteristics in their structure. Under this assumption, unique network features should be leveraged to classify an unknown network. Geng et al. (2012) conducted a study of biological network classification based on attribute vectors generated from global topological and label features. They discovered that networks from similar classes have similar characteristics, and network characteristics carry distinctions that can be leveraged in classification algorithms. Geng et al. (2012) found their feature-based classification models produced similar accuracy rates with less computational requirements than conventional kernel methods of measuring similarity between networks based on shared patterns.

Canning et al. (2018) investigated the use of network features for classification of complete real-world observed networks. Their research found that networks from differing classes do contain distinguishing structural features useful in network classification. Research prior to this study was mainly focused on classification of only synthetic networks or distinguishing networks within one specific class type. Canning et al. (2018) included synthetically generated networks among the real-world networks and discovered their classification model could identify the synthetic networks from real networks with great confidence. Their multiclass classification model using RF was successful in classifying both real-world and synthetic complete networks using only their network features.

These studies of feature-based classification presume complete network information in their methods. In contrast, we seek to examine a RF model that classifies a network as it is observed – even while incomplete.

2.2.4 Study of Incomplete Networks

Incomplete data is a reality of analyzing real-world networks. Portions of the observed data may remain unknown for different reasons such as data obstruction by excessive noise, non-respondent survey answers, deliberate concealment, or inaccessibility for observation (Garcia-Laencina et al. 2010). The proper handling of incomplete data is a critical requirement for accurate classification. An inapt approach can cause significant errors in classification results.

(Garcia-Laencina et al. 2010) discusses the following common techniques for analyzing incomplete data:

- **Exclusion** – deletion of incomplete datasets to analyze only completely observed data
- **Weighting** – modifying design weights to adjust for non-respondent data
- **Imputation** – an estimation of unobserved data is generated from known data features
- **Model-Based** – broad methods for modeling and making inferences based on data distribution or likelihood

Other emerging approaches for handling incomplete data include the use of ML techniques such as support vector machines (SVM), decision trees, and neural networks (NN) (Garcia-Laencina et al. 2010). However, when using any of these methods, we must be attentive to potential incidents of significant bias, added variance, or risks of generalizing estimated data (Garcia-Laencina et al. 2010).

Thus, we seek to develop a method for classifying an incomplete network without estimations to complete the network. Once classified, the methods of predicting unknown data can be customized to consider that network class's known properties, not just its observed features.

2.3 Definitions

In this thesis, "graph" and "network" are used interchangeably, making no distinction between the two. Explicitly, each are defined as follows,

Definition 2.3.1 *Graph*

> *A graph is "an ordered pair of disjoint sets (V, E) such that E is a subset of the set $V^{(2)}$ of unordered pairs of V... The set V is the set of vertices and E is the set of edges... An edge (x, y) is said to join the vertices x and y and is denoted by xy. Thus xy and yx mean exactly the same edge" (Bollobas 1998, p.1-2).*
>
> *Nodes and edges of a graph can contain additional information, "such as names or strengths, to capture more details of the system" (Newman 2010, p.2).*
>
> *A graph consists of "a set of nodes (vertices) and a set of edges (arcs) whose elements are pairs of nodes" (Ahuja et al. 1993, p.24).*

Definition 2.3.2 *Network*

> *A network in its most elementary form is "a collection of points joined together in pairs" by connections (Newman 2010, p.1). "[It] is a simplified representation that reduces a system to an abstract structure capturing only the basics of connection" patterns(Newman 2010, p.1).*
>
> *"Many practical problems can be represented by graphs. Emphasizing their application to real-world systems, the term network is sometimes defined to mean a graph in which attributes (e.g., names) are associated with the nodes and/or edges" (Graph Theory 2019).*

Networks can be a simplistic means of representing connections, our study focuses on distinguishing between networks' basic underlying structures. Thus, network representation is sufficient.

We consider the following network structure features for analysis and classification in this thesis,

Definition 2.3.3 *Number of Nodes*

> *The number of nodes represents the "number of components" in a network. (Barabasi 2016, sec.2.2)*

Definition 2.3.4 *Number of Edges*

> *The number of edges represents the "total number of interactions between the nodes" of a network. (Barabasi 2016, sec.2.2)*

Definition 2.3.5 *Average Distance*

> *A network's average distance can be measured as the average shortest path between pairs of connected nodes in the network (Barabasi 2016, sec.2.8).*
>
> *From Newman (2010, p.238), with V as the set of nodes in the network, n as the number of nodes in the network, and d(s, t) as the shortest distance from s to t, it is represented mathematically as,*

$$\textit{Average Shortest Distance} = \sum_{s,t \in V} \frac{d(s,t)}{n(n-1)}$$

Definition 2.3.6 *Degree*

> *The degree of a node is the number of edges attached to it (Ahuja et al. 1993, p.25).*

Definition 2.3.7 *Triangles*

A network triangle consists of three nodes, each connected by edges to both of the other nodes (Newman 2010, p.154).

Definition 2.3.8 *Chromatic Number*

The number of colors required to color the nodes of a network coloring the in such a way that no two vertices connected by an edge have the same color is the chromatic number (Newman 2010, p.113).

Definition 2.3.9 *Density*

The density of a network (also known as connectance) is the portion of the possible connections that are actual existing edges in the network (Newman 2010, p.117).

$$Density = \frac{NumberofEdges}{NumberofPossibleEdges}$$

Definition 2.3.10 *Transitivity*

Transitivity is the tendency for two neighbors of a vertex to also be neighbors of one another and is measured by the fraction of overall existing triangles to the number of overall possible triangles in the network (Newman 2010, p.191).

$$\textit{Transitivity} = 3\frac{\textit{Number of Triangles}}{\textit{Number of Possible Triangles}}$$

Definition 2.3.11 *Degree Assortativity Coefficient*

Degree Assortativity Coefficient measures the tendency for nodes of similar degree values to be connected (Newman 2010, p.219).

Newman (2003) also refers to it as the Pearson correlation coefficient of degree between pairs of linked nodes.

Definition 2.3.12 *Average Clustering Coefficient*

> *The clustering coefficient for a node is defined as the fraction of existing triangles to all possible triangles through that node (Newman 2010, p.298). A network's average clustering coefficient is measured as the average clustering coefficient for all of its nodes.*

$$\textit{Node Clustering Coefficient} = \frac{\textit{Number of Triangles (by Node)}}{\textit{Number of Possible Triangles (by Node)}}$$

Newman (2010, p.276) presents adjacency lists as a simple means for storing network data. In this thesis, we use adjacency lists to store our network data.

Definition 2.3.13 *Adjacency List*

> *Connected nodes are known as adjacent nodes (Ahuja et al. 1993, p.34). A node adjacency list is a common format for the storage of networks. It contains the set of nodes adjacent to each node in the network.*

$$\textit{Adjacency List} = A(i) = \{j \in N|(i, j) \in A\}$$

Definition 2.3.14 *Degree Centrality*

> *Centrality "quantifies how important [nodes](or edges) are in a networked system" (Newman 2010, p.8). Degree centrality is centrality measured by the number of edges connected to a node (Newman 2010, p.159).*

Of specific interest is studying important nodes, defined in this thesis to be the most connected nodes of a network. Newman (2010) asserts degree centrality to be a useful metric in determining node importance. Sparrow (1991) argues for examining centrality due to "incompleteness in the criminal databases [being systematic], anything but random." We focus on studying central nodes to examine the potential concealment of an organization's most important elements, as applicable to classifying incomplete criminal and terrorist networks. Including centrality into our methodology creates a model intelligence analysts can use for classifying these incomplete illicit networks. The next chapter introduces our methodology.

THIS PAGE INTENTIONALLY LEFT BLANK

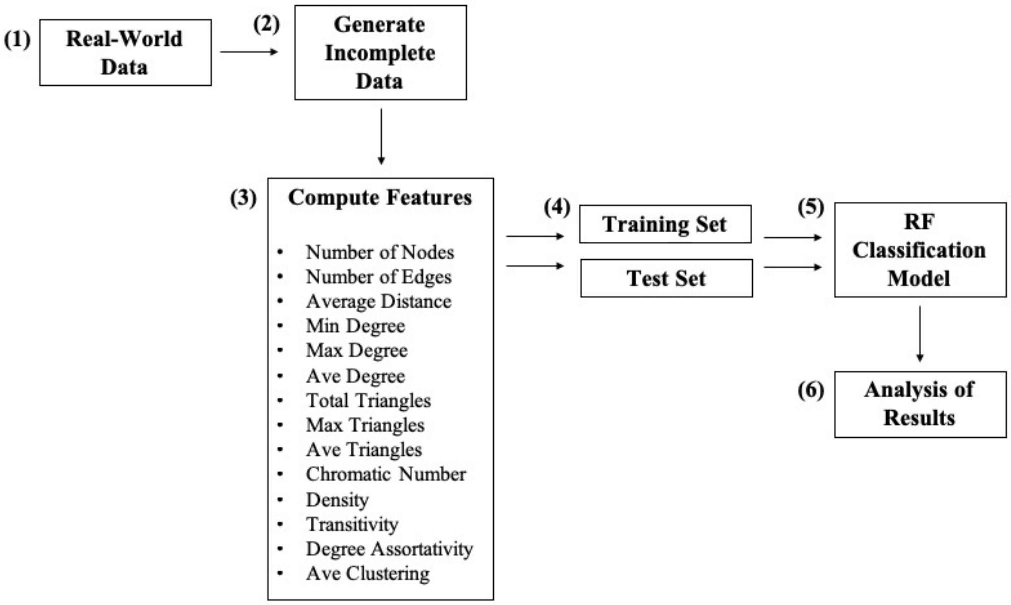
CHAPTER 3:
Model and Methodology

This chapter presents our methodology for classifying incomplete data and the network data selected for this thesis. The last two sections describe the classification model and discuss the required computational effort.

3.1 Methodology Overview

To examine effects of training a classification model with complete and incomplete information, we use the following approach illustrated in Figure 3.1.

Figure 3.1. Overview of Thesis Methodology



Our methodology is comprised of these six steps. In step 1, we obtain real-world data. In step 2, we generate incomplete data. In step 3, we compute network structural features. At step 4, we divide network data into training and test sets. In step 5 we train and test our RF model. Lastly, we analyze the results in step 6.

The steps of the method are:

1. Obtain complete real-world network data.
2. From the complete network, remove data to simulate an incomplete network.
3. Compute features for all complete and incomplete networks.
4. Build appropriate training and test sets.
5. Train RF classification model and test prediction.
6. Conduct analysis of results.

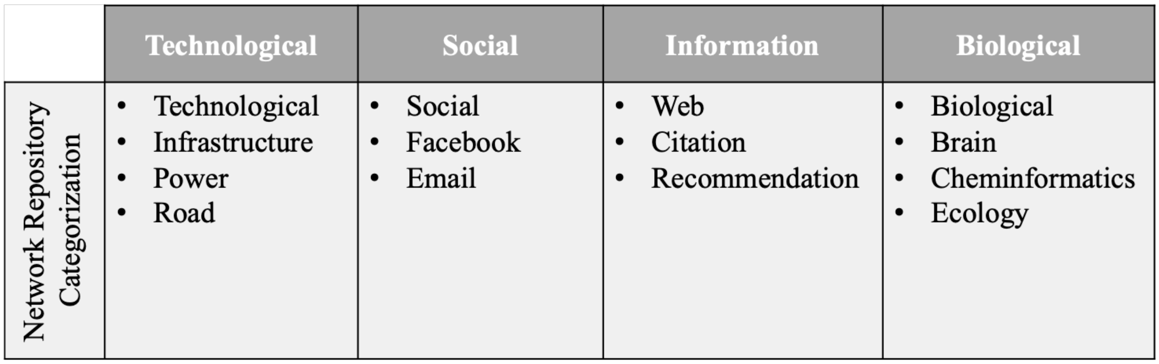
3.2 Network Data

In their study of network classification, Canning et al. (2018) found "[complete] synthetic graphs are trivial to classify as the classification model can predict with near-certainty the network model used to generate the synthetic graph." Our own preliminary exploration of incomplete synthetic networks, combined with Chia (2018) thesis findings, confirms similar results. Thus, this thesis only considers real-world observed networks for our classification model. We use the Rossi and Ahmed (2015) network repository as our primary data source for their comprehensive collection of network data.

3.2.1 Real-World Network Data

A sample of observed network data was obtained from the Rossi and Ahmed (2015) network repository. Our thesis examines networks in the following Rossi and Ahmed (2015) network repository categories: technological, infrastructure, power, road, social, Facebook, email, web, citation, recommendation, biological, brain, cheminformatics, and ecology. An exhaustive list of selected data is included in the appendix. For this thesis, we treat the data from Rossi and Ahmed (2015) network repository as complete networks and distribute them into the four common technological, social, information, and biological classes as described in Figure 3.2.

Figure 3.2. Networks Separated into Four Categories



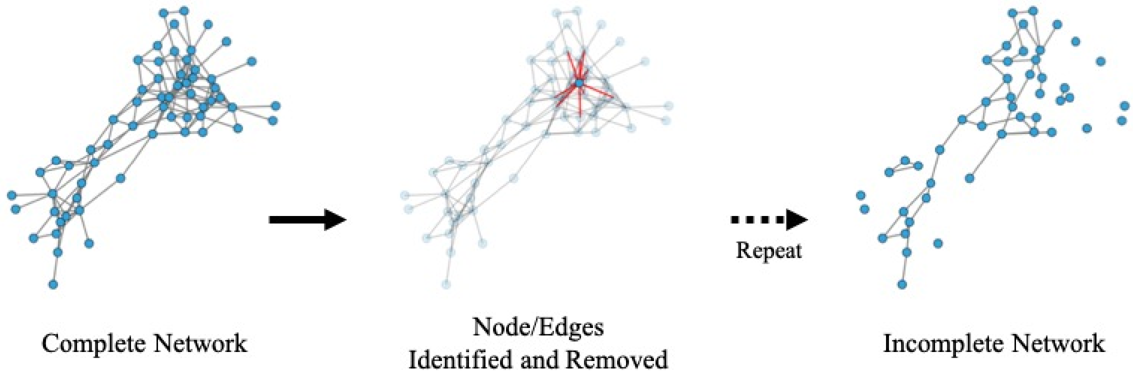
	Technological	Social	Information	Biological
Network Repository Categorization	• Technological • Infrastructure • Power • Road	• Social • Facebook • Email	• Web • Citation • Recommendation	• Biological • Brain • Cheminformatics • Ecology

Each network from the network repository was categorized by Rossi and Ahmed (2015). Using those categorizations, we group the networks into four general classes.

3.2.2 Generating Incomplete Data

From complete networks, we simulate incomplete network data by methodically removing nodes and edges in 5% increments. The general process is depicted in Figure 3.3. This generation of incomplete network data allows us to capture and study networks at varying stages of completeness.

Figure 3.3. Representing Incomplete Networks from Complete Networks



Beginning with a complete network, nodes and/or edges are selected and removed to simulate an incomplete network. This process is repeated to produce incomplete networks at differing levels of completeness, from 5% to 95% complete.

Starting with the original network (100% complete), we generate a sequence of incomplete networks, ranging from 5% to 95% complete. Thus, for a single network we have a range of representations at differing levels of completeness. For comparison, two techniques are developed to represent incompleteness, randomly and by degree centrality.

Incompleteness - Random

Two methods of generating incomplete data due to randomness are used.

1. Select nodes at random and remove. Remove all accompanying edges.
2. Select edges at random and remove. Nodes attached to any edge selected for removal remain in the network.

Incompleteness - Centrality

Two similar methods for generating incomplete data focused on centrality are also used.

1. Calculate degree centrality for each node. Identify the node with the greatest centrality and remove. Remove all accompanying edges.

2. Calculate degree centrality for each node. Remove edges attached to the central nodes. Nodes attached to any edge selected for removal remain in the network.

3.2.3 Network Features

In efforts to develop an efficient classification method, simple structural features are chosen to capture the characteristics of each network. The following features, as defined in Chapter 2, are computed using several algorithms in the Python library NetworkX (Hagberg et al. 2008), for all complete and incomplete networks.

- Number of Nodes
- Number of Edges
- Average Distance
- Minimum Degree
- Maximum Degree
- Average Degree
- Total Triangles
- Maximum Triangles
- Average Triangles
- Chromatic Number
- Density
- Transitivity
- Degree Assortativity Coefficient
- Average Clustering Coefficient

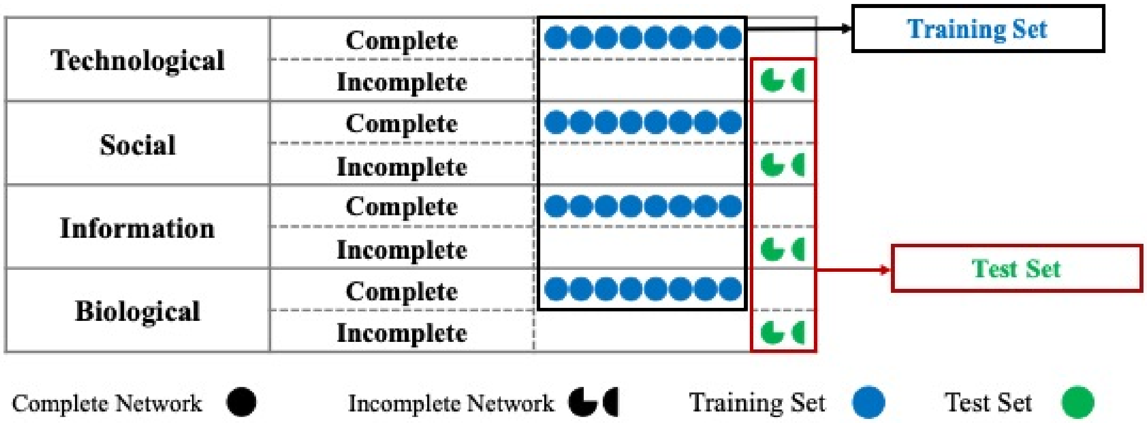
3.3 Model

Our supervised learning model consists of a training set, a test set, and a classifier.

3.3.1 Training Set

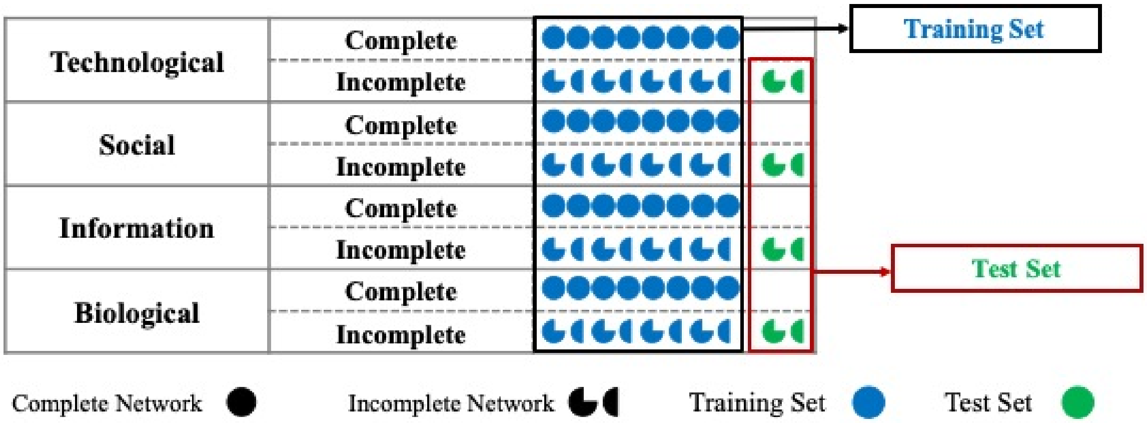
Ripley (1996, p.354) describes the training set as "a set of examples used for learning, that is to fit the parameters of the classifier". It is a subset of data used for training the model. For this thesis, two training sets are generated for comparison. The first training set follows the common method of training only with complete network data. Our approach incorporates incomplete network data into the training set. Figures 3.4 and 3.5 illustrate the two methods. This critical difference of allowing incomplete information into the training set provides the model with additional data to examine.

Figure 3.4. Common Method for Making Training and Test Sets



The blue denotes the subset of training set data. The green represents the test set data. In this method, only complete technological, social, information, and biological network data make up the training set. The test set contains incomplete networks at varying levels of completeness.

Figure 3.5. Proposed Method for Making Training and Test Sets



The blue denotes the subset of training set data. The green represents the test set data. In contrast to the common method, we include a subset of incomplete technological, social, information, and biological networks in the training set. The test set contains incomplete networks at varying levels of completeness.

3.3.2 Test Set

A test set is a subset of data reserved for assessing the trained model. Ripley (1996, p.354) describes the test set as “a set of examples used only to assess the performance of a fully specified classifier." This subset is not to be used in training the model but should be representative of the entire dataset. To assess the model’s ability to classify incomplete networks, we reserve a subset of incomplete network data for the test sets. As depicted by Figures 3.4 and 3.5 in green, we only include incomplete networks in the test set.

3.3.3 Random Forest Classifier

We use the `RandomForest` algorithm from the `sklearn.ensemble` Python module for predicting network class. The `RandomForest` algorithm is designed to follow Breiman (2001) standard RF methods and “fits a number of decision tree classifiers on various sub-samples of the dataset” (Pedregosa et al. 2011). The `sklearn.ensemble` Python module “combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class” after each decision tree of the collective forest is generated from a random sample of the training set (Pedregosa et al. 2011). This method leverages averaging to increase classification accuracy and control potential over-fitting (Pedregosa et al. 2011).

We choose the RF classification model for its accuracy, simplicity and efficiency. The `sklearn.ensemble` is well documented and maintained by experts in ML. We select the `sklearn.ensemble` `RandomForest` algorithm for its quality and ease of use.

3.4 Computational Effort

To process the complete networks from the Rossi and Ahmed (2015) network repository and create representations of incomplete networks, we use a standing desktop with twenty-four 3-GHz Intel(R) Xeon(R) CPU E5-2687W processors and 65 GB of memory. We process the complete networks using the following psuedocode.

```
– import network data
– convert to adjacency list for storage and reduction of data size
– find centrality for simulating centrality-based incomplete data
    – simulate incompleteness in 5% increments:  95%, 90%, ...5%
        * identify node or edge for removal
        * remove node and/or edge
        * save remaining network as adjacency list
    – calculate features listed in Section 3.2.3 for all networks
```

The processing time for each network is directly related to its network size. When generating incomplete networks using our centrality-based rule, the computational effort increases with the additional measure of each node's degree centrality. The processing time for simulating centrality-based incomplete networks directly relates to the overall density of each network.

We execute the RF model on a personal MacBook Air with a 1.3 GHz Intel Core i5 processor and 4 GB of memory. The model trained with only complete information requires on average 0.03 seconds. Executing the RF model on complete and incomplete information requires an average of 0.33 seconds to complete. This increase in computational time reflects the additional time needed to incorporate the incomplete network observations.

In the next chapter, we implement our algorithms and examine the effects of training the classification model with complete and incomplete network data. From our results, we analyze the effectiveness and potential applications for our model.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Results and Analysis

This chapter begins with an examination of the network features used in the classification model. Then, we analyze the model's classification accuracy. Lastly, Section 4.3 discusses potential applications for our incomplete network classification model.

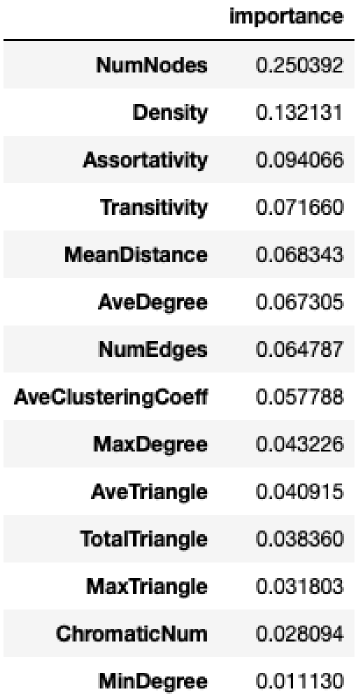
4.1 Feature Importance

The 14 network features from Section 3.2.3 are calculated for all networks in order to capture each of their observed structural characteristics. From the characteristics of training set networks, the model learns how to recognize other incomplete networks belonging to the same class. The model's predictive ability relies on a network feature's accurate portrayal of their network's characteristics. Thus, we conduct an evaluation of these features using the `sklearn feature_importances_` function and Shapley Additive Explanations (SHAP) values to examine their importance and contribution to the model.

4.1.1 Sklearn Feature Importance

The `sklearn feature_importances_` function scores each topological feature by its mean decrease impurity, where impurity is a measure of how often a network would be incorrectly classified if it was randomly classified according to the distribution of classes (Pedregosa et al. 2011). The mean decrease impurity is "defined as the total decrease in node impurity, weighted by the probability of reaching that node, averaged over all trees of the ensemble (Pedregosa et al. 2011)." Figure 4.1 displays the features' importance score. The number of nodes, density, and assortativity emerge as the three most important features, however, do not significantly distinguish themselves from the others.

Figure 4.1. Feature Importance by Mean Decrease Impurity



	importance
NumNodes	0.250392
Density	0.132131
Assortativity	0.094066
Transitivity	0.071660
MeanDistance	0.068343
AveDegree	0.067305
NumEdges	0.064787
AveClusteringCoeff	0.057788
MaxDegree	0.043226
AveTriangle	0.040915
TotalTriangle	0.038360
MaxTriangle	0.031803
ChromaticNum	0.028094
MinDegree	0.011130

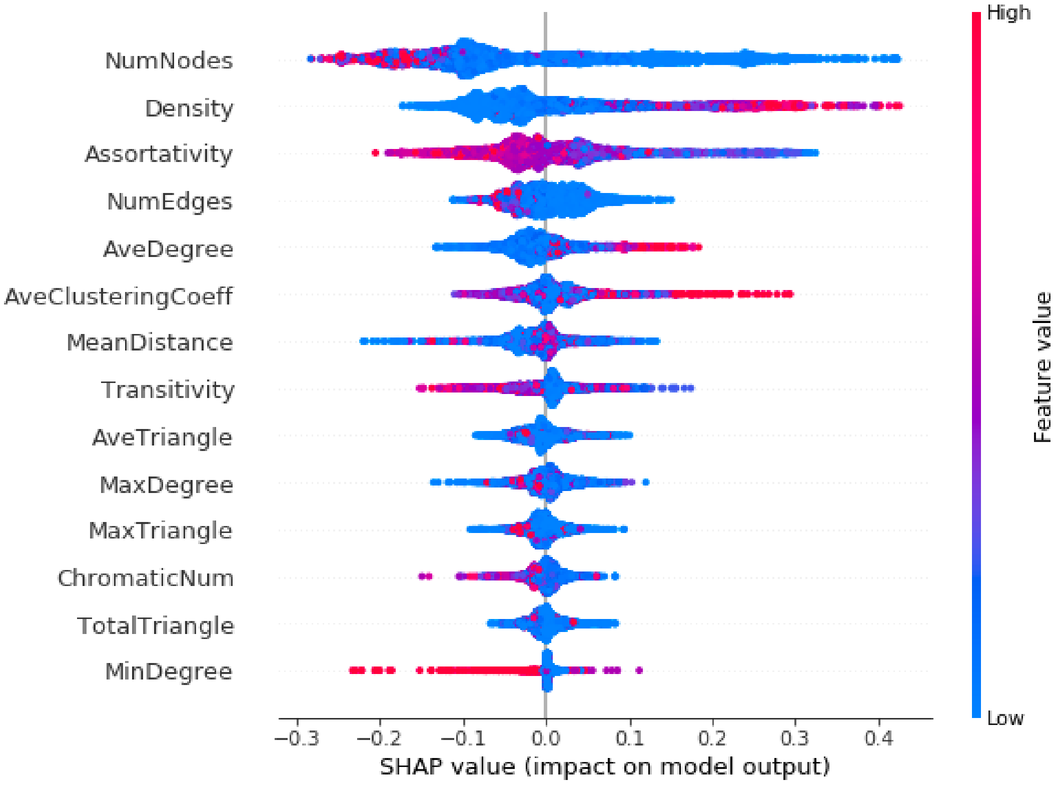
The ranking of features by mean decrease impurity importance score is intended to reveal features with the greatest impact to the model. While the number of nodes and network density have the greatest influence on our model, this ranking shows that their scores are not significantly greater than any other features' importance score. Also, all features have a relatively low importance value. Thus, we conclude none of the features are independently influential.

4.1.2 Shapley Additive Explanations

Shapley Additive Explanations (SHAP) merge game theory with six different methods for calculating feature importance (Lundberg and Lee 2017). The SHAP values are calculated to "represent a feature's responsibility for a change in the model output" to assist in the interpretation of a model's prediction output (Lundberg and Lee 2017).

Figure 4.2 shows the summary of all topological features and their SHAP values. Each network instance is plotted as a colored circle. Its color corresponds to the value of the feature as depicted in the legend. The x-axis characterizes the importance of the feature by its impact on the model, its SHAP value. This representation of feature importance reveals more than simply a score of importance. A low number of nodes, a high density, and a low assortativity have the greatest influence on the predictive model output. Interestingly, a high average clustering coefficient also impacts the model's prediction.

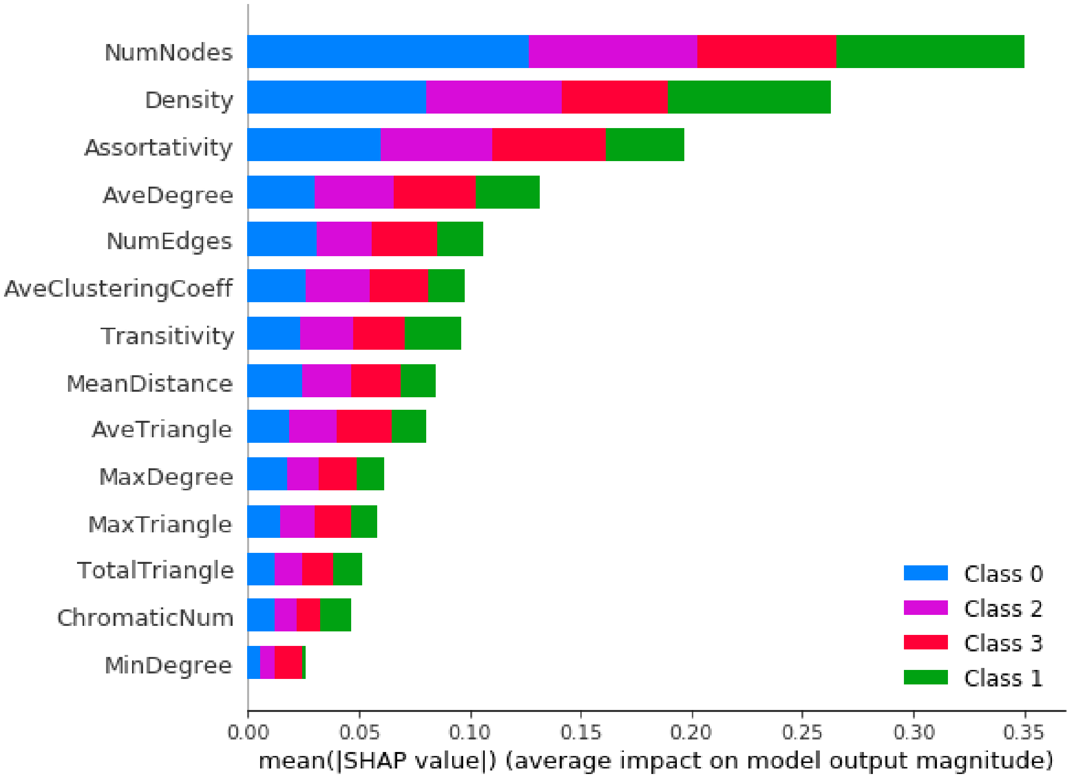
Figure 4.2. Feature Importance by SHAP Value



Each network occurrence is plotted in a color corresponding to the value of the feature. The x-axis characterizes the network feature's impact on the model. This representation of feature importance reveals low number of nodes, high density, and low assortativity have the most influence on our classification model.

Another method to explore feature importance by SHAP value is seen in in Figure 4.3. It depicts the average of SHAP values grouped by class type. This representation shows how each feature might impact the prediction of each class. Overall, the number of nodes, density, and assortativity emerge as the three most influential features for each class.

Figure 4.3. Feature Importance by SHAP Value



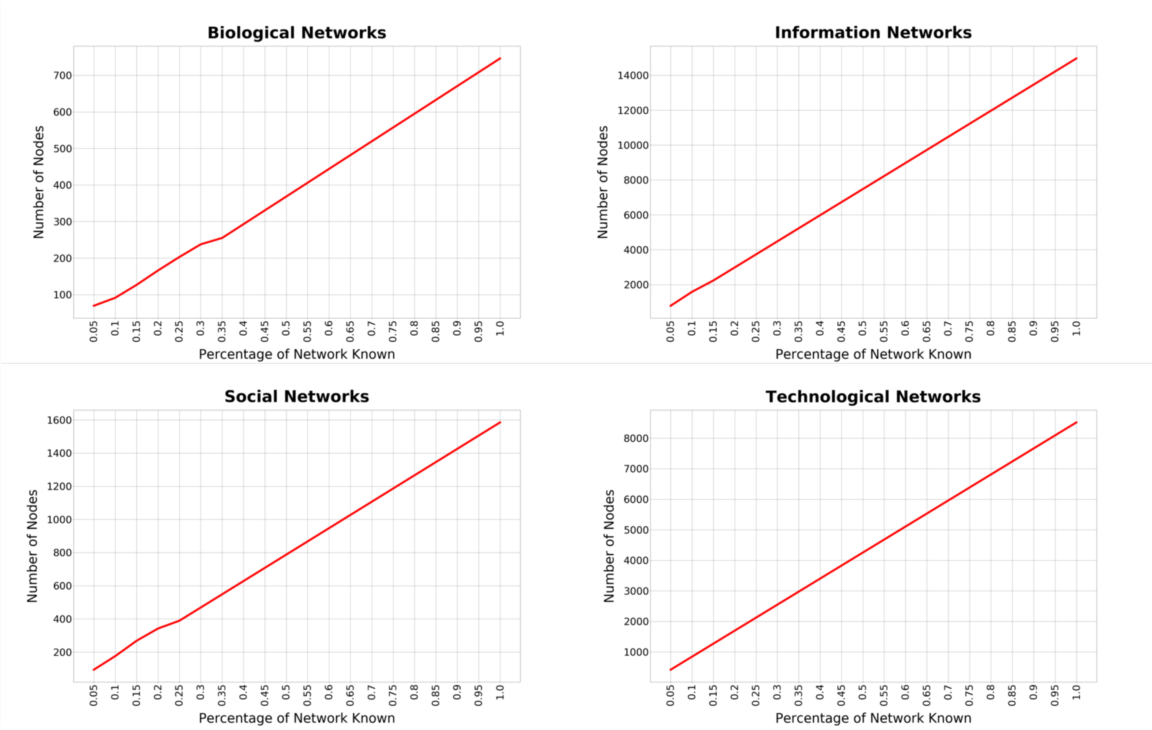
SHAP values are grouped and plotted by class type to show how each feature might impact the prediction of that class. Class 0 is biological networks. Class 1 is technological networks. Class 2 is social networks. Class 3 is information networks. The number of nodes, density, and assortativity are the three most influential features for each class.

4.1.3 General Behavior of Important Features

The number of nodes, density, and assortativity are identified as the three most influential features. To examine the general behavior of these three features, we inspect the changes in those features as the percentage of the known network changes. These changes are examined by class type.

Figure 4.4 shows the number of nodes in a network is directly related to the amount of information known about that network for all class types.

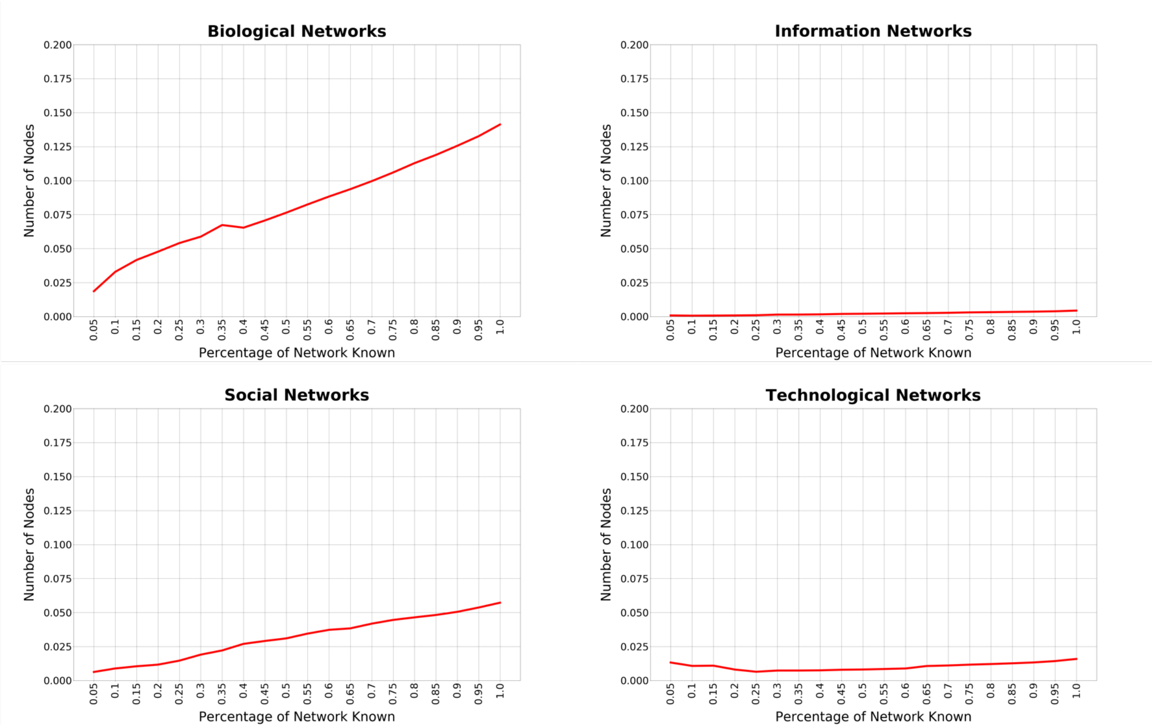
Figure 4.4. Average Number of Nodes by Class



The x-axis denotes the number of nodes. The y-axis represents the percentage of network known. In general, the number of nodes increases directly in proportion to the amount of network known for all classes.

As Figure 4.5 depicts, density responds differently for each of the class types. For technological and information networks, the density remains steady as the amount of known network information changes. In social networks, the density increases slightly with the increase in network information. The density in biological networks grows the most directly with the increase in network information.

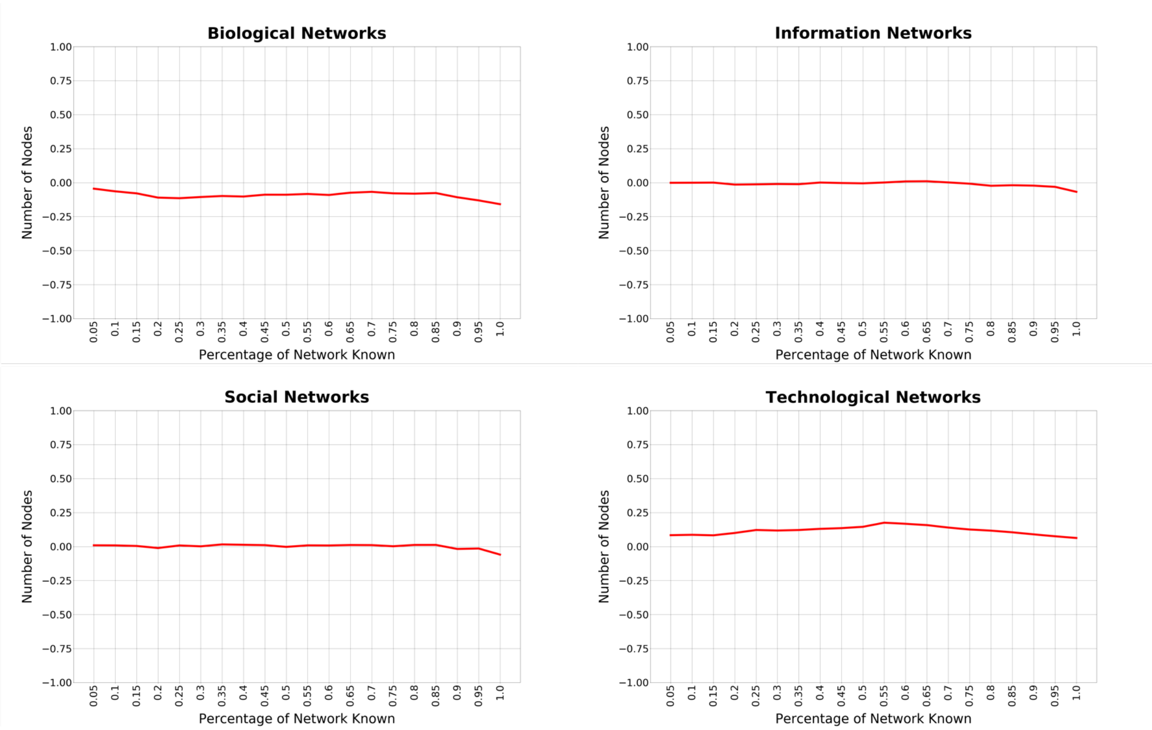
Figure 4.5. Average Network Density by Class



The x-axis denotes the network density. Its limits are from 0 to 0.2. The y-axis represents the percentage of network known. Average network density changes differently for each class as the known network information changes.

In Figure 4.6 we examine the degree assortativity remaining steady throughout changes in the amount of network information known. Overall, social and technological networks maintain a slightly positive assortativity on average. Whereas, biological networks have a slightly negative average assortativity. Information networks preserve an average neutral assortativity.

Figure 4.6. Average Degree Assortativity Coefficient by Class



The x-axis denotes the network density. Its limits are from -1 to 1. The y-axis represents the percentage of network known. In general, all classes maintain a steady neutral average degree assortativity coefficient as the amount of known network changes.

These observations confirm the limited influence any one of features has on predicting network class. As the network importance measures indicate, none of the features distinguish themselves as predominantly significant. With the complex subtleties and interdependencies of the features, this task of incomplete network classification is best suited for a ML method. This also validates our selection of a RF model.

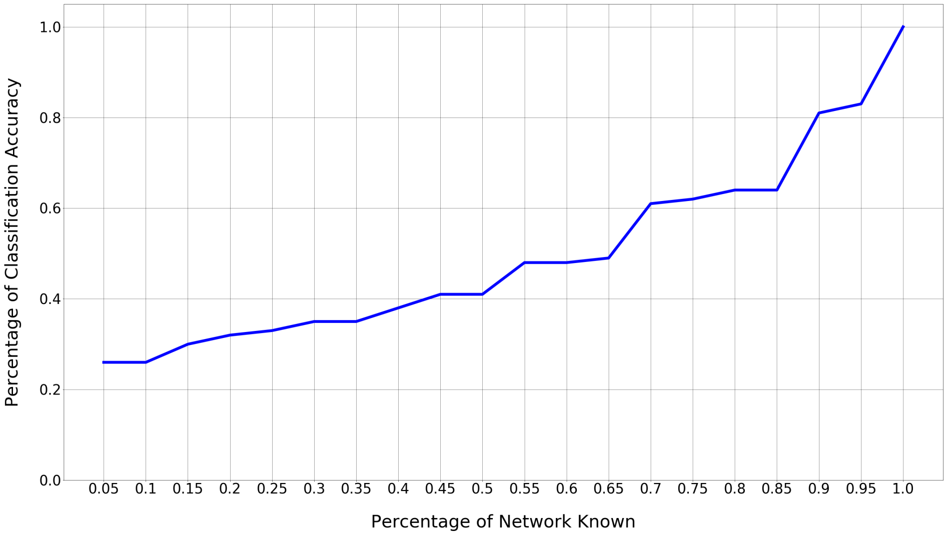
4.2 Random Forest Classification Accuracy

When considering the accuracy of RF classification techniques, we consider several cases, defined by the data used for training the model.

4.2.1 Trained on Only Complete Information

Figure 4.7 represents the classification accuracy achieved by the common method RF model trained by only complete network information. Beginning at 5% of the network known, the model has a classification rate of only 26%. Logically, as additional network information becomes known, the classification accuracy gradually increases. Of note, 90% of the network must be known to secure a classification accuracy greater than 80%. When all 100% of the network information is known, the model achieves near-certain classification accuracy.

Figure 4.7. Classification Accuracy - Trained on Complete Only

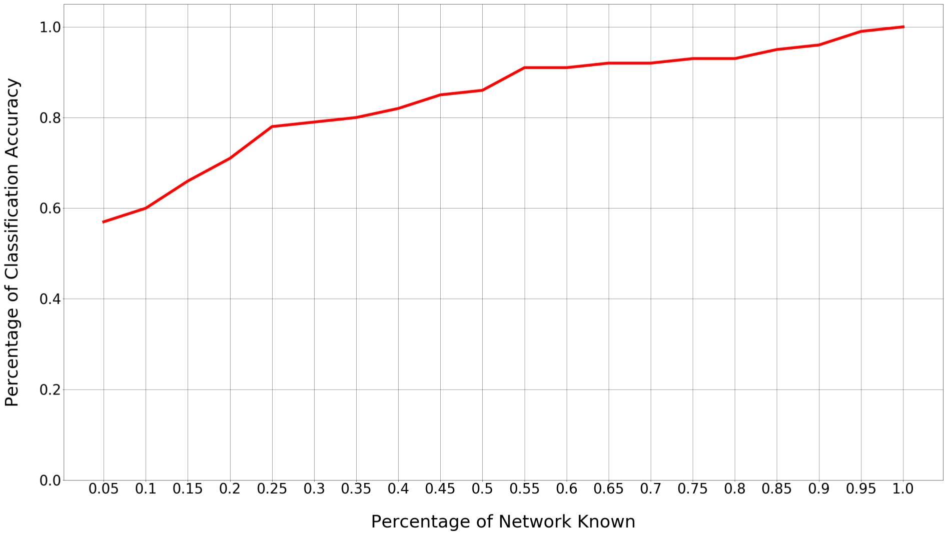


As the amount of network information grows, the classification accuracy increases. Still, the common method of training by only complete information requires 90% of the network to achieve a classification accuracy above 80%.

4.2.2 Trained on Complete and Incomplete Information

Figure 4.8 represents the classification accuracy achieved by our RF model trained by both complete and incomplete network information. There is a sharp classification accuracy increase from 5% to 25% of the network known. From 25% and greater, the accuracy steadily increases with the amount of network known. The lowest classification rate begins at 57% accuracy with merely 5% of the network known. At 25% of the network, we reach 78% classification accuracy. When all 100% of the network information is known, our model also achieves near-certain classification accuracy.

Figure 4.8. Classification Accuracy - Trained on Complete and Incomplete

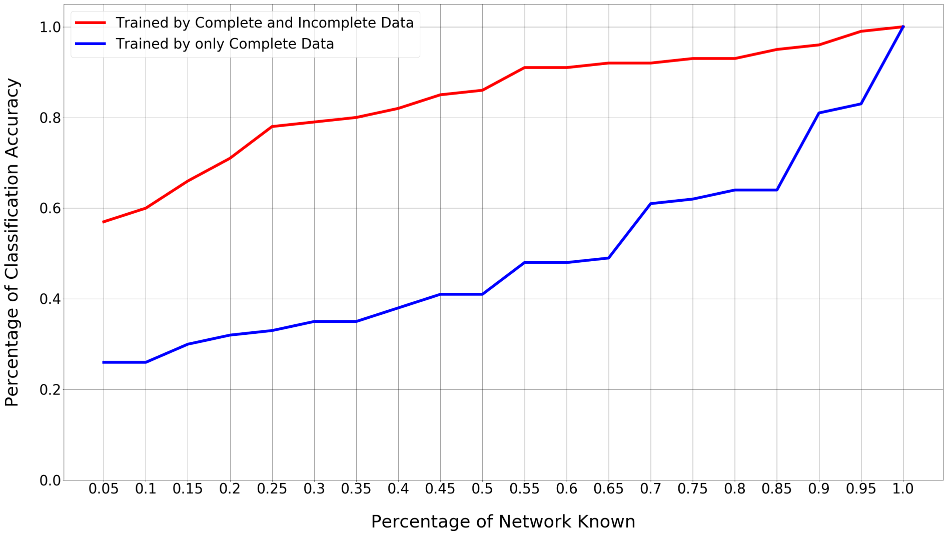


Our method of training with both complete and incomplete information achieves improves classification rates at all stages of network incompleteness. A minimum of 35% of the network reaches a classification accuracy over 80%.

4.2.3 Comparison of Both Random Forest Training Methods

We plot both RF model results together for a comparative analysis; Figure 4.9 depicts the classification accuracy for the two methods. At no point does the common method of training on only complete network data match the performance of our method which trains on both complete and incomplete data. To achieve an 80% accuracy, only 35% of the network information is necessary using a classification model trained on both complete and incomplete network information. Whereas, a model trained on only complete network information requires a minimum of 90% of the network information to be known. The stark difference between the two methods demonstrates the necessity of incorporating incomplete networks into the classification model training set.

Figure 4.9. Classification Accuracy - Training Method Comparison



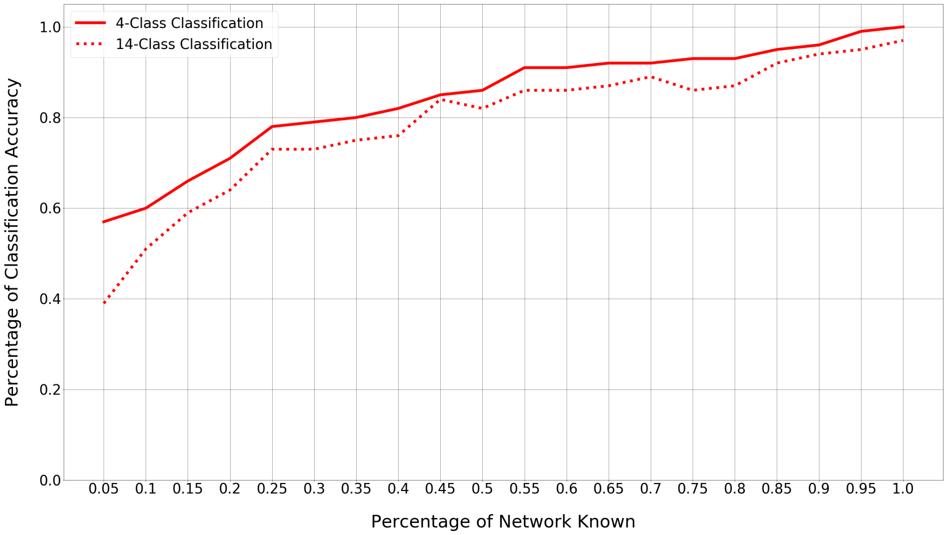
The difference between the two methods' accuracy rates emphasizes the benefits of training our RF classification model with both complete and incomplete information. To achieve an 80% classification accuracy, the common method of training on only complete data needs at least 90% of the network, while our method of training on complete and incomplete data only needs to know 35% of the network.

4.2.4 Classification by Network Category

To explore the potential for classification into more than four general classes, we also examine the performance of our model in predicting classification into 14 different network categories. The categories are technological, infrastructure, power, road, social, Facebook, email, web, citation, recommendation, biological, brain, cheminformatics, and ecology. We maintain the classification categories of each network as defined by the data source, Rossi and Ahmed (2015) network repository, for this test.

Figure 4.10 reveals a similarly trending, though slightly lower, accuracy compared to the four-class (technological, social, information and biological) classification. Classifying network information into 14 classes should be more difficult than into only four classes. However, our classification model trained with complete and incomplete information remains capable of classifying the networks at a similar accuracy rate. This result confirms the robustness of our methodology.

Figure 4.10. Classification Accuracy - Class vs Category



When expanding the classification to 14 network categories, our model is still capable of achieving accuracy rates comparable to the four-class classification. To achieve an 80% accuracy rate, only 45% of the network information is required.

4.3 Potential Direct Applications

The problem of incomplete information and need for classifying incomplete networks exists in multiple disciplines. For the Department of Defense (DoD) and the military, application in the intelligence community is most apparent. As intelligence about a network is collected, our classification method can be applied to classify the incomplete network without delay. An accurate classification prediction allows for action to be taken on limited information sooner. This classification aids intelligence analysts in their follow-on analysis of the network and contributes to the strategies of network dismantlement or influence, as the mission requires.

A network's classification helps determine how best to organize follow-on efforts to collect additional network information. With the network class known, principles of that class can aid in the prediction of future network connections or estimate the potential network structure development. This predictive ability helps intelligence analysts know "where to look" for additional network structure. Subsequently, political campaigning strategies, marketing tactics, and epidemiology studies can similarly benefit from this predictive growth analysis.

An example of another application in non-military disciplines includes epidemiology. The spread of a newly formed or mutated disease can be initially tracked and represented as a network. Using information about previously studied epidemics, the new disease network, while only partially known, can be assigned a class prediction that could aid medical professionals in their initial strategy for treatment and containment.

In the final chapter, we summarize our results and present potential areas for continuing analysis and future work.

CHAPTER 5:
Conclusion

This chapter summarizes our findings and analysis. Then, we introduce potential areas for continued future work.

5.1 Summary

In this thesis, we consider a method for classification of incomplete networks that includes incomplete network information in the training set. From real-world complete network information, we represent incomplete networks to allow the model to learn from the structural features of an incomplete network. We compare our method to the standard classification model that uses only complete network information in the training set.

Our results strongly indicate the need to include incomplete network representations in training the classification model. Incorporating incomplete networks at various stages of completeness allow the machine to examine and learn the nuances of incomplete networks. By allowing the machine to study incomplete network structural features, it has an improved ability to recognize and classify other incomplete networks. The RF Classification model requires minimal computational effort and can accomplish an efficient classification. We also confirm these simple, easily calculated network features are sufficient to classify an incomplete network.

5.2 Future Work

This thesis establishes a foundation for the continued study of incomplete networks. Our method of incomplete network classification provides preliminary insight into the benefits of incorporating incomplete network representation into training the model itself. However, the incomplete networks in our study are networks rendered incomplete due to only random or centrality-based reasons. Further exploration should be performed to consider other methods of representing incomplete networks to closely resemble realistic cases of incomplete information. Also, the networks we examine are limited to static observed networks

The expansion of study to include dynamic networks is recommended. Also, methods of accurately classifying sub-portions of a too-large network ought to be considered.

Additionally, whilst our research confirms the advantages of a standard supervised learning method in classifying incomplete networks, a deep learning approach ought to be considered to harness its capability and flexibility to process larger amounts of raw data through its incremental layered learning (LeCun et al. 2015). With the growth of accessible real-world data, large amounts of information will be available to train classification models. Though potentially time consuming to train, a deep learning method should have a faster classification speed and increased accuracy, especially as the known training set grows (LeCun et al. 2015).

Presently, the application of this classification method for a Department of Defense (DoD) Unmanned Autonomous Vehicle (UAV) network control project is being explored in a joint effort between the Operations Analysis (OA), Mechanical and Aerospace Engineering (MAE), and Computer Science (CS) Departments at the Naval Postgraduate School (NPS). The objective will be to use our classification model for quickly identifying when a degraded UAV network can no longer be classified as an operational mobile communication network. While a UAV network carries the features of a robust communication network, established principles, such as connectivity measures, can be applied in the tactical employment of the UAV network. The intent for this future study includes an incorporation of our classification model with an optimization model to rapidly predict and recommend specific vehicle tasks to sustain a robustly connected UAV network.

Combined with the foundation established in this study, these future research efforts allow for an enhanced understanding of incomplete networks and how to classify them. Efforts to incorporate this classification model in real applications is necessary for testing the model's practical implementation.

APPENDIX: Network Datasets

The primary data source for this thesis is from the Rossi and Ahmed (2015) network repository. The specific network datasets we use are listed in this appendix.

Technological Networks			
Infrastructure	Power	Road	Technological
inf-euroroad	power-494-bus	road-chesapeake	tech-as-caida2007
inf-openflights	power-662-bus	road-euroroad	tech-internet-as
inf-power	power-685-bus	road-minnesota	tech-p2p-gnutella
inf-USAir97	power-1138-bus		tech-pgp
	power-bcspwr09		tech-routers-rf
	power-bcspwr10		tech-WHOIS
	power-eris1176		
	power-US-Grid		

Social Networks		
Social	Facebook	Email
soc-advogato	socfb-Caltech36	email-dnc-corecipient
soc-ANU-residence	socfb-Haverford76	email-univ
soc-dolphins	socfb-nips-ego	email-enron-only
soc-firm-hi-tech	socfb-Oberlin44	
soc-hamsterster	socfb-Reed98	
soc-karate	socfb-Simmons81	
soc-physicians	socfb-Smith60	
soc-tribes	socfb-Swarthmore42	
soc-wiki-Vote	socfb-USFCA72	
	socfb-Wellesley22	

Information Networks		
Web	Collaboration	Recommendation
web-edu	ca-CondMat	rec-amazon
web-EPA	ca-CSphd	rec-movielens-tag-movies-10m
web-indochina	ca-Erdos992	rec-movielens-user-movies-10m
web-polblogs-2004	ca-GrQc	rec-yelp-user-business
web-spam	ca-netscience	
web-webbase-2001	ca-sandi_auths	

Biological Networks			
Biological	Brain	Cheminformatics	Ecology
bio-CE-GN	bn-cat-mixed-species_brain_1	ENZYMES_g1	eco-everglades
bio-CE-GT	bn-fly-drosophila_medulla_1	ENZYMES_g10	eco-florida
bio-CE-HT	bn-macaque-rhesus_brain_1	ENZYMES_g13	eco-foodweb-baydry
bio-CE-LC	bn-macaque-rhesus_brain_2	ENZYMES_g14	eco-foodweb-baywet
bio-celegans-dir	bn-macaque-rhesus_cerebral-cortex_1	ENZYMES_g15	eco-mangwet
bio-diseasome	bn-macaque-rhesus_interareal-cortical-network_2	ENZYMES_g16	eco-stmarks
bio-dmela	bn-mouse_brain_1	ENZYMES_g18	
bio-WormNet-v3-benchmark	bn-mouse_visual-cortex_1	ENZYMES_g101	
bio-yeast-protein-inter	bn-mouse_visual-cortex_2	ENZYMES_g102	
bio-yeast	bn-mouse-kasthuri_graph_v4	ENZYMES_g103	

List of References

Ahuja RK, Magnanti TL, Orlin JB (1993) *Network Flows: Theory, Algorithms, and Applications* (Prentice-Hall, New Jersey).

Alpaydin E (2014) *Introduction to Machine Learning* (Massachusetts Institute of Technology Press).

Barabasi AL (2016) *Network Science* (Cambridge University Press).

Bollobas B (1998) *Modern Graph Theory* (Springer, Heidelberg).

Breiman L (2001) Random forests. *Mach. Learn.* 45(1), https://doi.org/10.1023/A:1010933404324.

Canning JP, Ingram EE, Nowak-Wolff S, Ortiz AM, Ahmed NK, Rossi RA, Schmitt KRB, Soundarajan S (2018) Predicting graph categories from structural properties. *CoRR* 1805.02682, http://arxiv.org/abs/1805.02682.

Chia P (2018) Assessing the robustness of graph statistics for network analysis under incomplete information. Master's thesis, Department of Operations Analysis, Naval Postgraduate School, Monterey, CA, https://calhoun.nps.edu/handle/10945/58284.

Garcia-Laencina PJ, Sancho-Gomez JL, Figueiras-Vidal AR (2010) Pattern classification with missing data: a review. *Neural Comp. & App.* 19(2), https://link.springer.com/article/10.1007/s00521-009-0295-6.

Geng L, Semerci M, Yener B, Zaki MJ (2012) Effective graph classification based on topological and label attributes. *Stat. Analysis. & Data Mining* 5(4):265–283.

Graph Theory (2019) *Wikipedia*. Accessed March 14, 2019. https://en.wikipedia.org/wiki/Graph_theory.

Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using networkx. *7th Python in Sci. Conf.* (SciPy, Pasadena, CA), http://conference.scipy.org/proceedings/SciPy2008/paper_2.

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436–444.

Little R, Rubin D (2014) *Statistical Analysis with Missing Data*. http://ebookcentral.proquest.com/lib/ebook-nps/detail.action?docID=1775204.

Louppe G (2014) Understanding random forests: From theory to practice. Phd dissertation, Department of Electrical Engineering & Computer Science, University of Liège, Belgium, https://arxiv.org/pdf/1407.7502.pdf.

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. *Adv. in Neural Info. Pro. Systems 30* (NIPS 2017, Long Beach, CA), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mitchell T (1997) *Machine Learning* (McGraw-Hill, New York).

Murphy KP (2012) *Machine Learning: A Probabilistic Perspective* (Massachusetts Institute of Technology Press).

Newman M (2003) The structure and function of complex networks. *SIAM Review* 45(2):167–256.

Newman M (2010) *Networks: An Introduction* (Oxford University Press).

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Mach. Learn. Research* 12(Oct):2825–2830.

Ripley BD (1996) *Pattern Recognition and Neural Networks* (Cambridge University Press).

Rossi RA, Ahmed NK (2015) The network data repository with interactive graph analytics and visualization. *29th AAAI Conf. on A.I.* (AAAI15, Austin, TX), http://ryanrossi.com/pubs/aaai15-nr.pdf.

Sparrow MK (1991) The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks* 13(3):251–274.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California